3 Problem 1.3

Problem Statement 3.1. A source S has the digits $\{1, 2, 3, 4, 5, 6, 7\}$ as its alphabet. It sends out a digit sequence equally likely under the rule that $d_i < d_{i+1}$ except for when $d_i = 7$ where we allow that $d_i \geq d_{i+1}$ for all $i \geq 1$. Compute the entropy of this source.

3.1 Solution: Computing the entropy of the source

The *number of partitions of n into at most k parts*, denoted by $p_k(n)$ for $1 \leq k \leq n$, is defined by

$$p_k(n) = \sum_{\substack{j_1+2j_2+\cdots+kj_k=n \\ j_1,\ldots,j_k \geq 0}} 1.$$

We observe a fact from number theory (or mathematical partition theory) that

$$p_k(n) \sim \frac{n^{k-1}}{k! \cdot (k-1)!} \iff p_k(n) = \frac{n^{k-1}}{k! \cdot (k-1)!} + o_k(n^{k-1}), \text{ as } n \to \infty.$$

Next, let $\mathcal{D}$ denote the set of all possible valid digit sequences (of length at least one) from the source. For $n \geq 1$, set $\mathcal{D}_n := \mathcal{D} \cap \{1, 2, \ldots, 7\}^n$. That is, $\mathcal{D}_n$ denotes the set of valid digit sequences from the source of length exactly n.

Idea: We will bound $|\mathcal{D}_n|$ from above by a function depending on $p_7(n)$ for sufficiently large n. Each of the length-n digit sequences in $\mathcal{D}_n$ are equally likely as provided by the definition of the source. This then provides a lower bound on $\mathbb{P}[d^{(n)}]$ when $d^{(n)} \in \mathcal{D}_n$ is selected uniformly at random. Hence, we can use this bound in conjunction with the *asymptotic equi-partition principle* (AEP) to show that as $n \to \infty$,

$$H(S) \sim -\max_{d^{(n)} \in \mathcal{D}_n} \frac{\log_2 \mathbb{P}[d^{(n)}]}{n}.$$

Then because the right-hand-side of the previous equation tends to zero when evaluated with a lower bound derived from the probability based on the probability estimate involving $p_7(n)$, we can conclude that $H(S)$ converges to the remaining constant term in our bounds.

Lemma 3.1. *If the source generates a length-n string $d^{(n)} = d_1 d_2 \cdots d_n$ for $n \geq 1$, then for any $m \in \{1, 2, \ldots, 7\}$*

$$\mathbb{P}[d_1 = m] = \frac{1}{7},$$

and if $2 \leq i \leq n$, then

$$\mathbb{P}[d_i = m] = \begin{cases} \frac{20}{363}, & m = 1; \\ \frac{70}{1089}, & m = 2; \\ \frac{28}{363}, & m = 3; \\ \frac{35}{363}, & m = 4; \\ \frac{140}{1089}, & m = 5; \\ \frac{70}{363}, & m = 6; \\ \frac{140}{363}, & m = 7; \\ 0, & \text{otherwise.} \end{cases}$$

Proof. The case for d_1 is clear as we equally likely select the first digit in the string. For $i \geq 2$, we have a recursion given by

$$\mathbb{P}[d_i = m < 7] = \frac{1}{7}\mathbb{P}[d_{i-1} = 7] + \sum_{k=1}^{m-1} \frac{1}{7-k} \cdot \mathbb{P}[d_{i-1} = k]$$

$$\mathbb{P}[d_i = 7] = 1 - \sum_{k=1}^{6} \mathbb{P}[d_i = k].$$

Subject to the condition that the sum over all of the probabilities must equal one at each i, the last two equations imply the numerical probabilities we have claimed above. Note also that for $2 \leq m < 7$ and $2 \leq i \leq n$

$$\mathbb{P}[d_i = m] = \left(1 + \frac{1}{8-m}\right) \mathbb{P}[d_i = m-1].$$

This last observation leads to the conjecture stated below in more generality. $\square$

Proposition 3.1. *For all sufficiently large n, and any $d^{(n)} \in \mathcal{D}_n$, we have that*

$$\frac{1}{\mathbb{P}[d^{(n)}]} \leq 7e^{210} \cdot n^{21} \cdot p_7(n) \cdot 7^n \sim e^{210} \cdot \frac{n^{21}}{6!^2} \cdot 7 \cdot \left(\frac{363}{140}\right)^{n-1}.$$

Proof. Given any valid digit sequence $d^{(n)} \in \mathcal{D}_n$, the strictly increasing substrings of the sequence can be broken down into individual sequences of bounded sizes in $\{1, 2, \ldots, 7\}$. So for any length-n string generated by the source, it can be decomposed into an equivalent representation by a finite sequence of substrings within these bounded lengths. Hence, we can bound the size of $\mathcal{D}_n$ for any sufficiently large n as follows:

$$\begin{aligned}
|\mathcal{D}_n| \leq & \sum_{\substack{k_1+2k_2+\cdots+7k_7=n \\ k_1,\ldots,k_7 \geq 0}} \#\,(\text{ways to arrange the } k_1+\cdots+k_7 \text{ substring lengths}) \times \\
& \qquad \times \prod_{i=1}^{7} \#\,(\text{ways to write a valid sequence of length } i)^{k_i} \\
& \leq \sum_{\substack{k_1+2k_2+\cdots+7k_7=n \\ k_1,\ldots,k_7 \geq 0}} \frac{n^{1+2+\cdots+7}}{k_1!k_2!\cdots k_7!} \times \prod_{i=1}^{7} \left(\binom{7}{i} \cdot i!\right)^{k_i} \\
& \leq n^{21} \times \sum_{\substack{k_1+2k_2+\cdots+7k_7=n \\ k_1,\ldots,k_7 \geq 0}} \frac{210^{k_1+k_2+\cdots+k_6}}{k_1!\cdots k_7!} \\
& \leq n^{21} \cdot p_7(n) \times \left(\sum_{k=0}^{n} \frac{210^k}{k!}\right)^7 \leq e^{7\times 210} \cdot n^{21} \cdot p_7(n) \sim \frac{e^{7\times 210} \cdot n^{21}}{7! \cdot 6!}.
\end{aligned}$$ $\square$

An immediate corollary of the proposition is that for large n and $\varepsilon \to 0^+$, the probability of a digit sequence $d^{(n)} \in \mathcal{D}_n$ satisfying the AEP satisfies

$$\mathbb{P}[d^{(n)}] \geq \frac{1}{7e^{7\times 210} \cdot n^{21} \cdot p_7(n) \cdot \left(\frac{363}{140}\right)^{1-n}}.$$

Similarly, we can see that $|\mathcal{D}_n| \geq 1$ by taking the string consisting of n consecutive 7 digits. Hence, $d^{(n)} \in \mathcal{D}_n$

$$\frac{1}{\mathbb{P}[d^{(n)}]} \geq 7 \cdot \left(\frac{363}{140}\right)^{n-1}.$$

We find that for large n (as $n \to \infty$)

$$\log_2\left(\frac{363}{140}\right) + o(1) \leq -\frac{\log_2 \mathbb{P}[d^{(n)}]}{n} \leq \log_2\left(\frac{363}{140}\right) + \frac{\log_2\left(\frac{7e^{7\times 210}}{7!\cdot 6!}\right) - \log_2\left(\frac{363}{140}\right) + 21\log_2(n)}{n} + o(1).$$

So since the right and left-hand-sides above become zero-values except for the constant terms as $n \to \infty$, we conclude by the AEP that $H(S) = \log_2\left(\frac{363}{140}\right) \approx 1.37454$.

Conjecture 3.1 (Generalization)**.** *In general, if we repeat the construction in the problem, except that the alphabet for the source $\mathcal{S}_k$ has digits $\{1, 2, \ldots, k\}$ for $k > 1$, we expect that $H(\mathcal{S}_k) = \log_2(H_k)$ where $H_k = \sum_{1 \leq j \leq k} j^{-1}$ is a* first-order harmonic number*. For example, in the lecture 3 example where $k := 9$, we expect that $H(\mathcal{S}_9) = \log_2\left(\frac{7129}{2520}\right) \approx 1.50028$. Moreover, for large k, as $k \to \infty$, we expect that $H(\mathcal{S}_k) = \frac{\log\log k}{\log 2} + O\left(\frac{1}{\log k}\right)$.*